Humboldt-Universität zu Berlin
Mathematisch-Naturwissenschaftliche Fakultät
Institut für Informatik

Expose, Robert Kühnen:

Automatic Identification a nd Classification of Tennis Strokes During Professional Tennis Matches using an Ensemble of TSC techniques

Gutachter/innen: Dr. Patrick Schäfer
Tony Bagnall (University of Southampton)

1 Introduction

The majority of research into detecting and classifying tennis strokes has primarily centered on sensor-based technologies, which employ various types of sensors attached to equipment or players to capture detailed biomechanical data. These sensor-based approaches are designed to be player-independent, allowing for broad applicability across different athletes without the need for customization. While these methods are prevalent, there is also a significant body of research exploring visual-based approaches. Although less common, these visual methods leverage the rich cues provided by video data to analyze stroke mechanics. Like sensor-based techniques, visual methods generally aim to be player-independent, functioning effectively across a diverse range of players. However, they face inherent limitations such as their dependency on the camera's field of view, which can restrict their ability to detect or classify strokes performed out of sight. In contrast, sound-based approaches, which have been less explored, offer unique advantages by identifying strokes through the auditory data generated during play, unhindered by visual obstructions. This auditory capability is crucial, especially since no current methodologies effectively utilize the trajectory of the ball for classification purposes without visual input. To date, limited research has ventured into the realm of sound-based classification of tennis strokes, with one notable attempt employing Mel Frequency Cepstral Coefficients (MFCCs) and Hidden Markov Models to achieve this objective [2]. However, the landscape of machine learning has evolved significantly in recent years, introducing new models and techniques that have the potential to revolutionize how we approach sound-based classification tasks. The comprehensive review and experimental evaluation by Middlehurst et al. [4], popularly known as 'Bake off redux,' provides an updated perspective on the field of time series classification (TSC) algorithms. Since the first major 'bake off' study, the field has seen significant advancements with the introduction of novel algorithms and approaches. This recent study revisits the state of TSC, evaluating the performance of new algorithms against established ones using an expanded set of datasets from the UCR archive. It further extends the taxonomy of TSC algorithms to include new categories like convolution and feature-based algorithms, as well as deep learning approaches. The study finds that algorithms such as Hydra+MultiROCKET [1] and HIVE-COTEv2 [3] demonstrate superior performance on both the current and newly introduced TSC problems. This research underscores the rapid evolution of TSC techniques and offers valuable insights for the application of these advancements in complex data scenarios like sound-based tennis stroke classification. This thesis seeks to explore these recent machine learning innovations to detect and classify tennis strokes purely from sound data.

By doing so, it aims to not only fill a significant gap in the current literature but also to test the applicability of these advanced techniques in a new domain. The successful application of these methods could pave the way for their use in other fields, thereby expanding the horizons of sound-based classification beyond the realm of sports.

2 Problem Description

2.1 Potential problems

In contrast to sensor-based and visual-based methods, a sound-based approach encounters unique challenges. Background noise is a prevalent issue, arising from various sources such as spectators, announcements, and commentators. Moreover, the sounds of player and equipment interactions with the court—such as sliding, the racket hitting the ground, or the ball striking obstacles like court surroundings or benches—add further complexity. The presence of other players engaging in matches nearby can also lead to acoustic confusion, complicating the model's ability to isolate and classify stroke sounds accurately.

Additionally, employing audio data from different sources introduces inconsistencies in audio quality. Variations in recording equipment, microphone placement, and environmental conditions such as wind or humidity can all influence the quality and consistency of the recorded sounds. Even the sounds of the strokes themselves can be deceptively similar, complicating the classification process further. Importantly, the direction of the stroke relative to the microphone significantly affects the frequency of captured sound due to the Doppler effect. This phenomenon adds a layer of complexity to our analysis as it influences how sound waves are perceived depending on whether the stroke is moving towards or away from the recording device.

Racket specifications such as string type, head size, weight, and the use of vibration dampers, although rarely used in professional matches, can affect the acoustic signature of strokes. Fortunately, the similarity across all rackets used by a professional player does mitigate this issue slightly. Another potential problem is players' grunting, which can overlap with the sound of the stroke being classified.

The typical use of only one microphone, often embedded within the camera, introduces additional variability. The varying distances between the sound source and the recording device can affect the amplitude and clarity of the recorded sounds. Additionally, class imbalances are an unavoidable challenge. While we have attempted to mitigate this by reducing the number of classes to eight, certain strokes like serves, forehands, and backhands are inherently more frequent than others.

While we are confident in our ability to accurately label strokes, it's important to acknowledge the inherent limitations imposed by the video frame rate. Videos from different sources may render at various frame rates, which can significantly affect our ability to precisely time the initiation of each stroke. As a result, there may be slight misalignments in the timing of stroke initiation, which, while minimal, could potentially influence the accuracy of our classifications.

These combined factors make sound-based tennis stroke classification not only an innovative challenge but also a complex problem requiring sophisticated solutions to effectively interpret and categorize the nuanced acoustic data.

2.2 Tennis shots

In tennis there are various shots. Unfortunately there isn't one widespread definition for them. Below a list of 22 shots and their descriptions from the International Tennis Federation (ITF). Table 1

For this study we use the following 8 classes for shots.

We define those shots as follows:

1. Serve - Shot that starts a rally
2. Smash - A shot hit with the ball above head height hit with a downward trajectory towards the opponent's court
3. Volley: Shot, that got hit mid-air before the ball bounces
4. Lob - A shot where the racket goes from low to high, resulting in the ball having A net clearance greater than 4m
5. Drop Shot - A softly hit shot that lands close to the net, often played with backspin
6. Slice - A shot that creates backspin on the ball, resulting in a lower bounce
7. Forehand - Shot, that got hit after bouncing once, where the palm of the racket hand faces towards the direction of the shot
8. Backhand - Shot, that got hit after bouncing once, where the back of the racket hand faces towards the direction of the shot (can be both one-handed or two-handed)

We further clarify the definitions with the flowchart below: Figure 1

Shot	**Shot Description**
Serve	The initial stroke to start each point, struck after tossing the ball into the air and aiming to land it in the diagonal service box.
Groundstroke	Played after the ball bounces, typically from near the baseline.
Volley	Hit before the ball bounces, usually near the net.
Forehand	Groundstroke or volley hit with the palm of the racket hand facing towards the shot.
Backhand	Groundstroke or volley hit with the back of the racket hand facing towards the shot.
Ace	A serve that the opponent cannot reach, resulting in a point.
Approach Shot	Played while moving toward the net.
Backspin	Stroke that causes the ball to spin backward.
Chip and Charge	A sliced approach shot followed by a move to the net.
Crosscourt	Shot that sends the ball diagonally across the court.
Deep	A shot that lands near the opponent's baseline.
Down the line	Shot hit close and parallel to the sideline.
Drop shot	A softly hit shot that lands close to the net with backspin.
Flat	Hit with little or no spin.
Half volley	Struck just after the ball bounces.
Kick serve	A serve with topspin that bounces high on the receiver's side.
Lob	Shot that passes over the opponent's head and lands in the court.
Passing Shot	Struck past an opponent at the net.
Slice	Shot with backspin causing a lower bounce.
Smash	A powerful overhead shot.
Topspin	Creates a forward rotation, causing a higher bounce.
Tweener	Struck between the legs, often when the player's back is to the net.

Table 1: Summary of Tennis Shots and Descriptions from `https://m.itftennis.com/en/about-us/organisation/tennis-glossary/`

Definition of newly introduced terms:

- Rally - Series of back-and-forth shots between players, starting with the serve and continuing until a point is scored
- Groundstroke - Shot hit after ball has bounced once on the court
- Net clearance - Refers to how high the ball goes over the net

Shot	Label
Serve	S
Smash	SM
Volley	V
Lob	L
Drop Shot	DS
Slice	SL
Forehand	FH
Backhand	BH

Table 2: Tennis Shots and their Labels

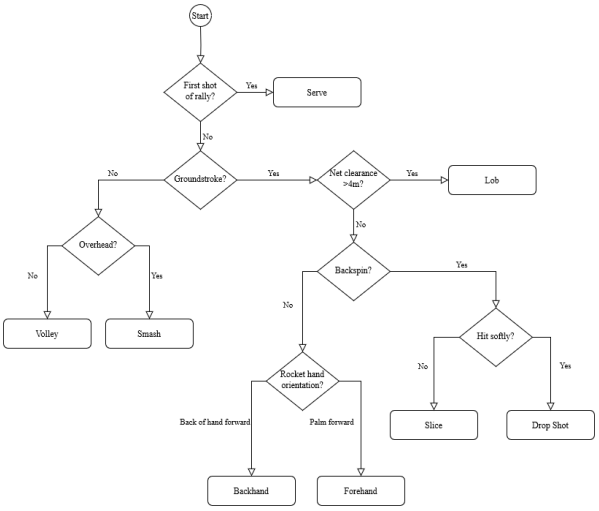

Figure 1: Flowchart to describe shots we defined

- Overhead - Shot hit above the head
- Backspin - A type of spin where the ball rotates backwards towards the player hitting it, causing it to bounce low after it lands on the opponent's side

3 Research Questions

We investigate the capabilities of our model through the following research questions:

RQ1: How well can our model distinguish between tennis strokes and other sounds occurring during tennis matches?

RQ2: How well do different techniques of the aeon toolkit, used for our ensemble, perform on their own?

RQ3: How accurately can our ensemble model classify the detected tennis strokes?

RQ4: How does our model perform when applied to data from an entirely unseen dataset?

RQ5: Is our model player dependent or a universal stroke model?

RQ6: Are our predictions fast enough to be implemented in real-time?

4 Benchmark Creation

For benchmark creation, we employ Kinovea [1], which is a software specifically designed to annotate critical moments within video footage. Each significant tennis stroke is marked with a 'key image'. Kinovea captures a snapshot at each key image and generates an accompanying comment. The title of the comment is automatically set to the timestamp of the key image. However, for the purposes of our study, this title is replaced with a label corresponding to the tennis stroke captured in the key frame, as delineated in Table 2. Kinovea's comment mechanism facilitates this labeling process, as shown in the example below Figure 2.

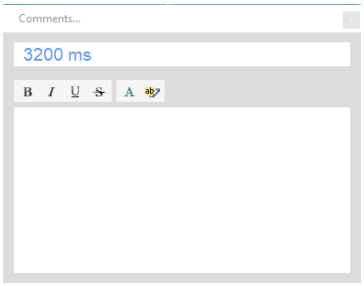

Figure 2: Example of a Kinovea comment created for a key image. The title field shows the timestamp "3200 ms", which we replace with the appropriate stroke label.

For more screenshots of the labeling process please refer to Figure 9 to Figure 13.

Upon completion of labeling, we export the data to an XML format. Parsing this XML with common Python libraries proves to be challenging. Therefore, we will utilize Python's 're' package to perform regular expression searches, effectively extracting the necessary label (Name) and timestamp (Time) information from the XML content, as illustrated below:

<!— *Additional XML content* —>
<Row>

[1] `https://www.kinovea.org/`

```
        <Cell ss:StyleID="header">
          <Data ss:Type="String">Name</Data>
        </Cell>
        <Cell ss:StyleID="header">
          <Data ss:Type="String">Time</Data>
        </Cell>
      </Row>
<!-- Additional XML content -->
```

This approach to data preparation ensures the integrity and accuracy of our benchmark, which is essential for the forthcoming phases of training and validating our machine learning models.

5 Methodology

The methodology for detecting and classifying tennis strokes from audio data encompasses several preprocessing and analysis stages. The initial stage involves acquiring the data, which could be downloaded from YouTube in video format. This video data is then converted into audio format, specifically into waveform audio file format (.wav), which is suitable for audio processing and analysis tasks. This is done using Python's moviepy.editor [2] package.

Subsequently, we address the stereo-to-mono conversion to standardize the audio data. Stereo recordings contain two separate audio channels, which may introduce unnecessary complexity for our analysis. By converting the stereo audio to mono, we ensure a singular, consistent audio channel, thereby simplifying the subsequent processing stages.

Instead of committing to specific techniques at this juncture, our approach will be experimental, involving a variety of techniques from the aeon toolkit. This toolkit provides a range of algorithms and methodologies that are suited to time series analysis and have the potential to be adapted to our audio classification task. We aim to iteratively test and evaluate these techniques to determine which ones are the most effective for classifying the distinct sound patterns associated with different tennis strokes.

Below 2 spectograms Figure 3 and Figure 4 as well as a waveform Figure 5 of a highlight video of a professional tennis match from YouTube. [3]

As our methodology evolves, we will document our findings and adapt our approach

[2] https://github.com/Zulko/moviepy

[3] https://www.youtube.com/watch?v=U5Af1jGgYqA

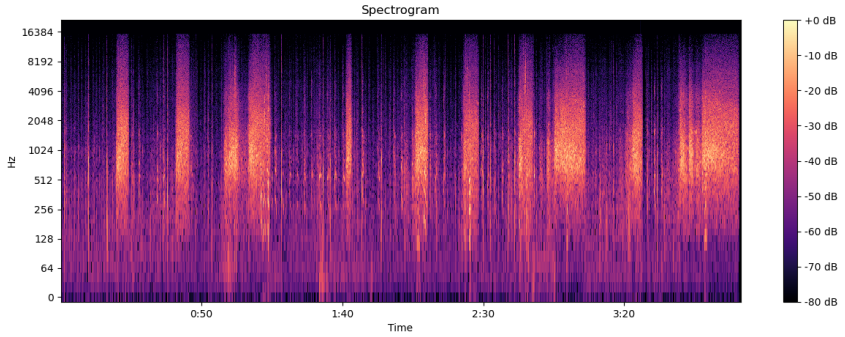

Figure 3: Spectrogram representation of audio data extracted from a tennis match video, illustrating the frequency distribution over time for 129 tennis strokes.

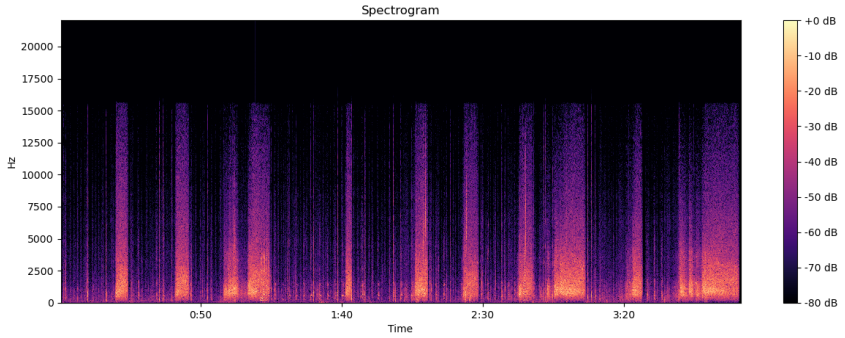

Figure 4: Spectrogram with a linear frequency scale displaying the same audio data as Figure 3. A notable frequency cutoff at 16 kHz is present, likely resulting from limitations in the microphone quality or processing techniques used prior to the video's upload or a compression algorithm, as this characteristic is not observed in other datasets we tested.

accordingly. The flexibility of the aeon toolkit will allow us to refine our analysis as we gain more insights into the data and the performance of the various models at our disposal.

As our research unfolds, the methodology section will evolve to document the iterative process of experimentation and discovery. The immediate goal is to demonstrate the viability of audio-based tennis stroke classification using the aeon toolkit, with the hope of laying the groundwork for a system that could eventually become an asset to sports analytics. Although achieving a fully robust and accurate classification system may not be an immediate outcome, the insights gained through this exploratory phase will contribute valuable knowledge to the field and may inform future developments in sound classification across various domains.

So far, all spectrogram and waveform visuals are derived from the audio of the same tennis match. Below are some visuals for a 14-stroke rally of length 18 seconds, from

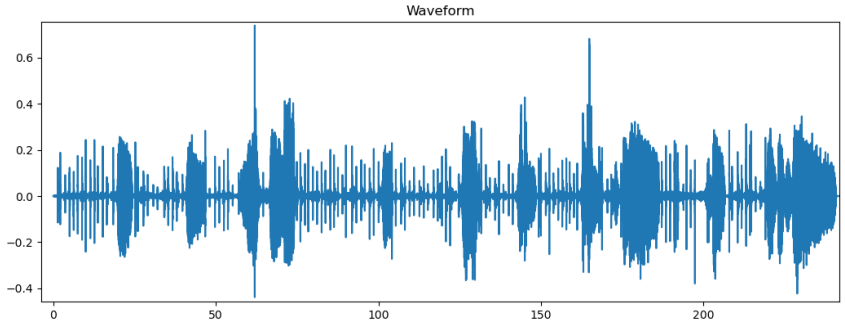

Figure 5: Waveform representation of the same audio data as Spectrograms above. Figure 3 and Figure 4

the very start of this match. This segment was chosen for its uninterrupted sequence of strokes.

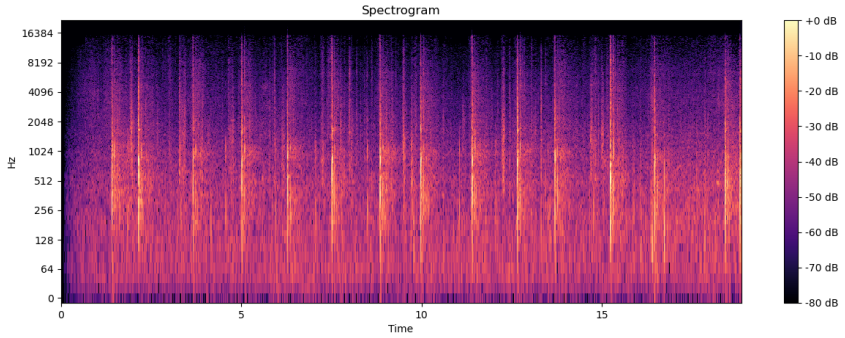

Figure 6: Spectogram for a 14-stroke rally from the very start of the match, for which the spectogram is shown in whole in Figure 3

6 Threats to Validity

This research acknowledges several potential threats to validity that could impact the outcomes and generalizability of the study.

6.1 Dataset Size and Quality

A primary concern is our labeled dataset being too small or the data quality being insufficient. Limited data can constrain the model's ability to learn and generalize, while poor quality data may lead to inaccurate modeling of tennis stroke sounds.

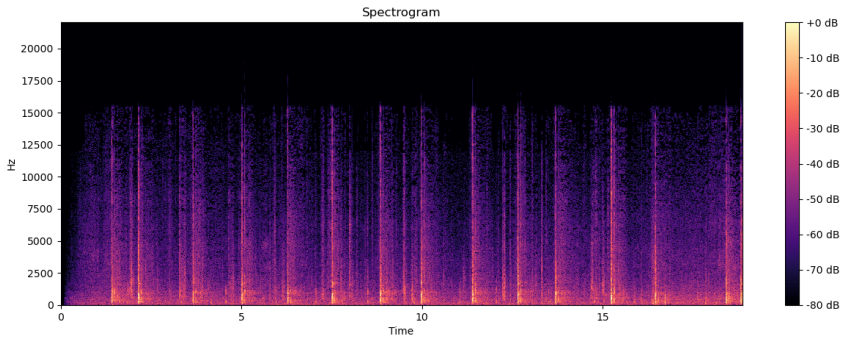

Figure 7: Spectogram with a linear frequency scale displaying the same audio data as Figure 6

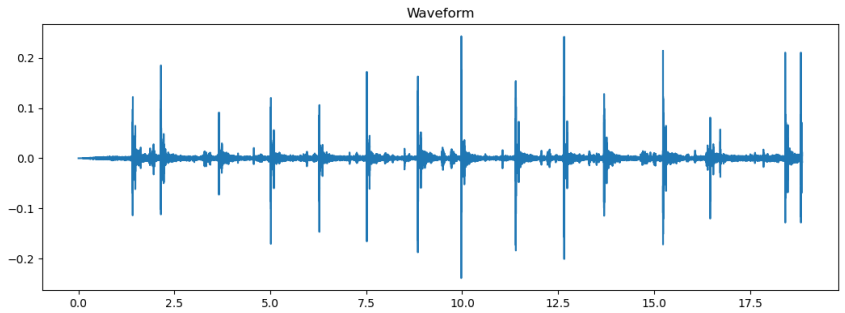

Figure 8: Waveform representation of the same audio data as Spectrograms above. (Figure 6 and Figure 7)

6.2 Labeling Accuracy

Labeling accuracy remains a critical concern in this study. Despite a decade of experience in tennis, which aids significantly in identifying and categorizing tennis strokes, the potential for human error cannot be entirely eliminated. More critical, however, are the inherent limitations posed by the use of Kinovea for video analysis. The software, while robust for visual tagging and review, renders videos at varying frame rates depending on the source material. This variability can significantly impact the precision with which stroke initiation times are marked. The frame rates typically range, leading to potential misalignments in the exact timing of stroke initiations. Such misalignments, even if minor, could influence the timing accuracy of labeled strokes, affecting the subsequent training and performance of the classification models. These factors combined could potentially introduce slight inaccuracies in the dataset, thus posing a challenge to the validity of the study's outcomes.

6.3 Researcher Expertise

The researcher's expertise in both tennis and machine learning may influence the study's direction and outcomes. While possessing extensive experience in tennis, the researcher does not claim to be an authority in the sport or in the field of machine learning. To counterbalance this, the study leverages established methodologies and consults domain experts for validation of the approaches used.

6.4 External Validity and Generalization

The study's external validity, particularly its generalizability across different datasets, is a crucial point of consideration. The choice of datasets can significantly influence how well the model performs in diverse real-world scenarios. Ensuring that the datasets encompass a wide range of playing conditions, player skill levels, and stroke variations is vital for enhancing the model's robustness and applicability to different tennis environments. However, our current dataset is derived from audio recordings of only one match, which introduces a limitation due to the high labor intensity required to accurately label each stroke. This constraint may result in the model being potentially player-dependent, as the variability across different players and matches is not represented. Consequently, while the model may perform well within the specific context of this match, its ability to generalize to other players or conditions might be restricted.

These threats to validity are recognized and addressed throughout the study to strengthen the reliability and applicability of the research outcomes.

References

[1] A. Dempster, D. F. Schmidt, and G. I. Webb. HYDRA: competing convolutional kernels for fast and accurate time series classification. *CoRR*, abs/2203.13652, 2022.

[2] G. Hunter, K. Zienowicz, and A. Shihab. The use of mel cepstral coefficients and markov models for the automatic identification, classification and sequence modelling of salient sound events occurring during tennis matches. *The Journal of the Acoustical Society of America*, 123:3431, 06 2008.

[3] M. Middlehurst, J. Large, M. Flynn, J. Lines, A. Bostrom, and A. J. Bagnall. HIVE-COTE 2.0: a new meta ensemble for time series classification. *Mach. Learn.*, 110(11):3211–3243, 2021.

[4] M. Middlehurst, P. Schäfer, and A. Bagnall. Bake off redux: a review and experimental evaluation of recent time series classification algorithms, 2023.

Appendix

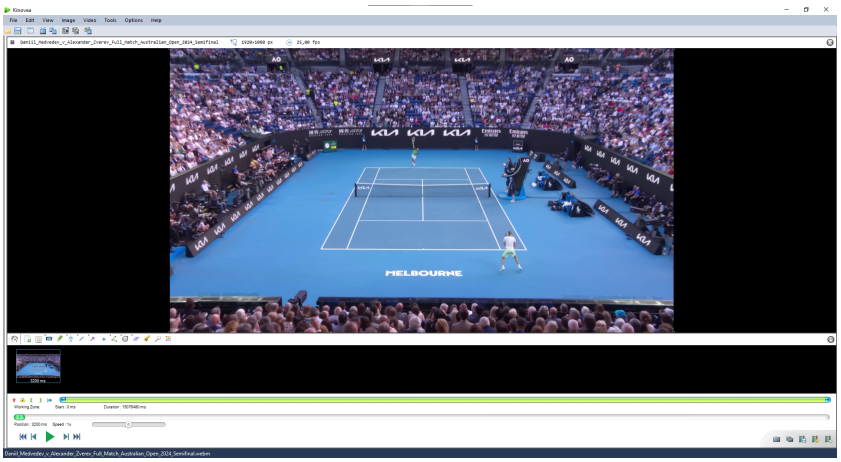

Figure 9: Screenshot of Kinovea after creating a key image

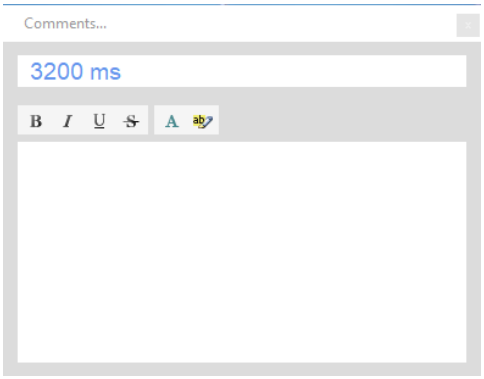

Figure 10: Screenshot of a Kinovea comment created for a key image. The title field shows the timestamp "3200 ms", which we replace with the appropriate stroke label. (same graphic as Figure 2)

Spectograms and Waveform for data from a different matches highlight video [4]:

[4] https://www.youtube.com/watch?v=Ts_jFq2gjnI

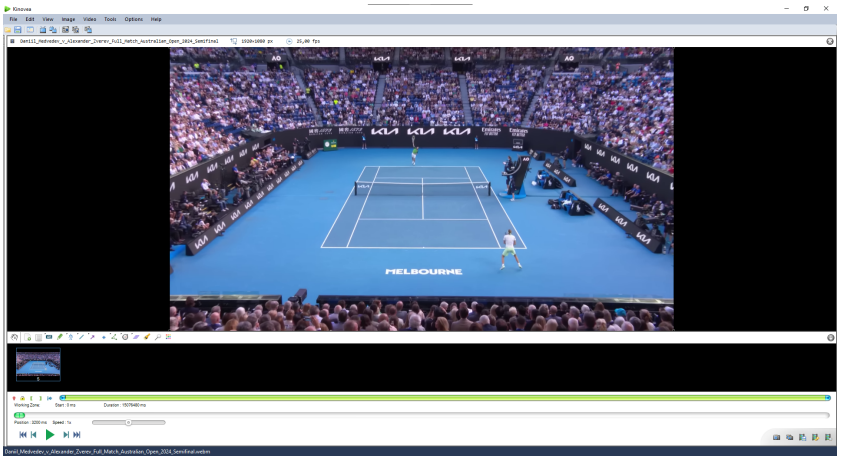

Figure 11: Screenshot of Kinovea after creating a key image and changing the resulting comments title

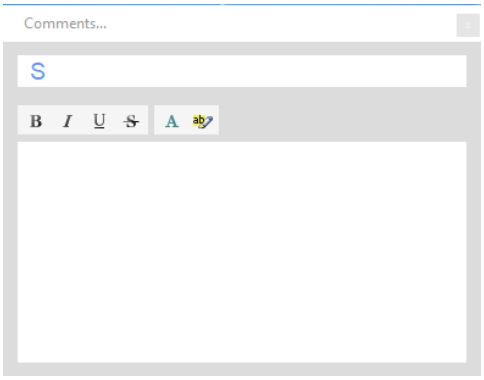

Figure 12: Screenshot of a Kinovea comment after changing its title to the stroke label

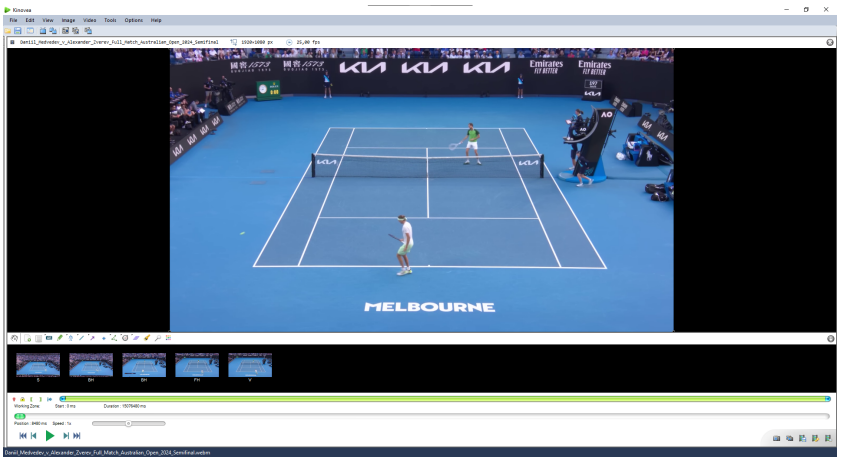

Figure 13: Screenshot of Kinovea after creating key images and changing the resulting comments titles for the first rally of the match.

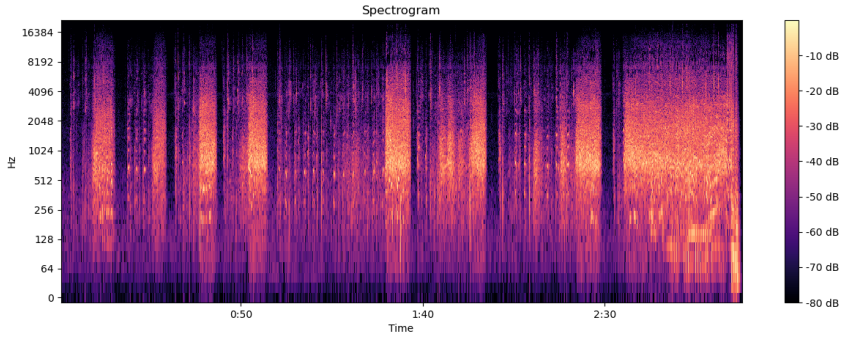

Figure 14: Spectrogram representation of audio data extracted from a tennis match highlight video, illustrating the frequency distribution over time

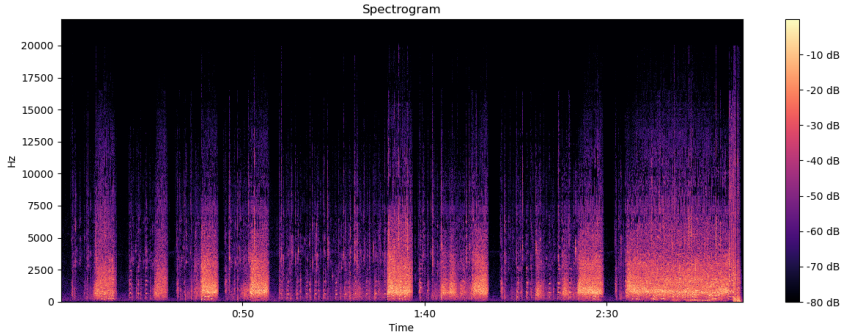

Figure 15: Spectrogram with a linear frequency scale displaying the same audio data as Figure 14

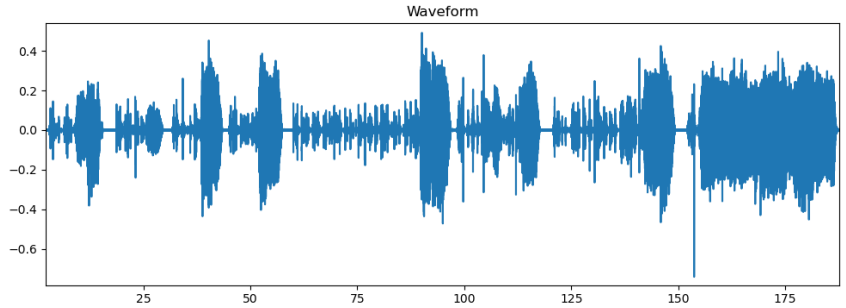

Figure 16: Waveform representation of the same audio data as Spectrograms above. Figure 14 and Figure 15